Theoretical assignment 1; 20 points total

Theoretical Deep Learning course, MIPT

Problem 1

4 points total.

Let $\mathcal{L}(W)$ be a scalar function of a matrix W. Its derivative is a matrix with following entries:

$$\left(\frac{d\mathcal{L}}{dW}\right)_{ij} = \frac{d\mathcal{L}}{dW_{ij}}.$$

Let y and x both be vectors. Let $\Sigma_{xx} = \mathbb{E}_{x,y\sim\mathcal{D}} xx^T$ and $\Sigma_{yx} = \mathbb{E}_{x,y\sim\mathcal{D}} yx^T$, where $\mathcal{D}$ is the data distribution.

1. **1 point.** Let

   $$\mathcal{L}(W) = \frac{1}{2}\mathbb{E}_{x,y\sim\mathcal{D}}\|y - Wx\|_2^2.$$

   Prove that

   $$\frac{d\mathcal{L}}{dW} = W\Sigma_{xx} - \Sigma_{yx}.$$

2. **1 point.** Let

   $$\mathcal{L}(W_1, W_2) = \frac{1}{2}\mathbb{E}_{x,y\sim\mathcal{D}}\|y - W_2W_1x\|_2^2.$$

   Prove that

   $$\frac{d\mathcal{L}}{dW_1} = W_2^T(W_2W_1\Sigma_{xx} - \Sigma_{yx});$$
   $$\frac{d\mathcal{L}}{dW_2} = (W_2W_1\Sigma_{xx} - \Sigma_{yx})W_1^T.$$

3. **1 point.** Let

   $$\mathcal{L}(W_{1:H}) = \frac{1}{2}\mathbb{E}_{x,y\sim\mathcal{D}}\|y - W_HW_{H-1}\ldots W_1x\|_2^2.$$

   Prove that

   $$\frac{d\mathcal{L}}{dW_k} = W_{k+1}^T\ldots W_H^T(W_H\ldots W_1\Sigma_{xx} - \Sigma_{yx})W_1^T\ldots W_{k-1}^T.$$

4. **1 point.** Let

$$\mathcal{L}(W_{1:H}) = \frac{1}{2}\|Y - W_H W_{H-1} \dots W_1 X\|_F^2,$$

where X and Y are matrices of dimensions $d_0 \times m$ and $d_H \times m$ respectively. Prove that

$$\frac{d\mathcal{L}}{dW_k} = W_{k+1}^T \dots W_H^T (W_H \dots W_1 X - Y) X^T W_1^T \dots W_{k-1}^T.$$

Problem 2

1 point.

Let $\mathcal{W}_{n,k,d}$ be the set of all $n \times k$ matrices of rank $\leq d$, and let $1 \leq n < k$. Prove that for any d such that $1 \leq d < n$ the set $\mathcal{W}_{n,k,d}$ is non-convex.

Problem 3

5 points total.

Consider a deep linear net:

$$\hat{y}(x, W_{1:H}) = W_H W_{H-1} \dots W_1 x.$$

Let d_i be the width of i-th layer (i.e. $W_i \in \mathbb{R}^{d_i \times d_{i-1}}$). The square loss of this net on a finite dataset $X \in \mathbb{R}^{d_0 \times m}$, $Y \in \mathbb{R}^{d_H \times m}$ is given as follows:

$$\mathcal{L}_{deep}(W_{1:H}) = \frac{1}{2}\|Y - W_H W_{H-1} \dots W_1 X\|_F^2.$$

Assume $H \geq 2$.

1. **1 point.** Prove that $W_{1:H} = 0$ is a critical point.
2. **3 points.** Assume $XY^T \neq 0$. Assume also, there are no bottlenecks in architecture, i.e. $\min_k d_k = \min\{d_0, d_H\}$. Using results of Lu & Kawaguchi (2017)[1] prove that $W_{1:H} = 0$ is a saddle point.
3. **1 points.** Prove that if $H \geq 3$ then then the hessian of $\mathcal{L}_{deep}$ at $W_{1:H} = 0$ is zero.

Problem 4

3 points.

Consider the same setting as in the previous problem.

[1] `https://arxiv.org/abs/1702.08580`

Assume $d_0 = \ldots = d_H$; hence all matrices W_k are square. Prove that if a local minimum $(W_1, \ldots, W_H)$ of $\mathcal{L}_{deep}$ is such that $W_H W_{H-1} \ldots W_1$ is of full rank, then it is a global minimum of $\mathcal{L}_{deep}$.

We had already proven a more general result at the lecture, however, our proof relied on the lemma about SVD of perturbed matrix (which we formulate quite informally and didn't prove). You are asked to prove the result for the special case of square matrices without relying on this lemma.

Problem 5

Up to 7 points.

All results in Lu & Kawaguchi (2017) are formulated for a finite dataset of size m. We can, however, assume an arbitrary data distribution $\mathcal{D}$:

$$\mathcal{L}_{deep}(W_{1:H}) = \frac{1}{2}\mathbb{E}_{x,y\sim\mathcal{D}}\|y - W_H W_{H-1} \ldots W_1 x\|_2^2,$$

$$\mathcal{L}_{sh}(R) = \frac{1}{2}\mathbb{E}_{x,y\sim\mathcal{D}}\|y - Rx\|_2^2,$$

where we assume correlation matrices $\Sigma_{xx} = \mathbb{E}(xx^T)$ and $\Sigma_{yx} = \mathbb{E}(yx^T)$ to be of full rank. Is it possible to generalize Theorem 2.1 (see paper) for this setting? If there could be problems, indicate where.

We expect an answer of the form: "A new setup doesn't affect Lemma 3.1. Proof of Lemma 3.2 requires slight modification: <what to modify>. This is not obvious, how to generalize Theorem 3.1, since <your arguments>". An answer, where necessary modifications which actually generalize Theorem 2.1 are given, will receive the highest degree.